Tarea Complementaria 2 Encoding

🎯 Objetivo

Comprender los principales **sistemas de codificación de caracteres (encoding)** utilizados en informática y programación:

ASCII, **ISO-8859** y **Unicode (UTF-8, UTF-16, UTF-32)**.

🔹 Código ASCII

📖 Definición

El **ASCII (American Standard Code for Information Interchange)** es el **primer estándar de codificación de texto** desarrollado en 1963 por la **ANSI (American National Standards Institute)**. Representa letras, números y símbolos mediante **valores numéricos entre 0 y 127**.

- Cada carácter se representa con **7 bits** (un byte menos un bit libre).
- Incluye:
  - Letras inglesas (`A–Z` , `a–z`)
  - Dígitos (`0–9`)
  - Signos de puntuación (`. , ; : ? !`)
  - Caracteres de control (como `\n` salto de línea, `\t` tabulación, etc.)

> 📘 *Ejemplo:*
> La letra **A** en ASCII → código **65 (01000001)**.

🔸 ASCII extendido

Posteriormente se amplió a **8 bits (0–255)**, añadiendo caracteres especiales, acentos y símbolos gráficos.

Sin embargo, **no era un estándar único**, y cada fabricante (IBM, Windows, DOS, etc.) lo implementó de forma diferente.

> 💡 Esto generó incompatibilidades entre sistemas.
> Para resolverlo, surgieron estándares internacionales como **ISO-8859**.

🔹 ISO-8859

📖 Definición

ISO-8859 es una **familia de codificaciones de 8 bits** desarrollada por la **ISO (International Organization for Standardization)**.

Cada versión cubre un conjunto de idiomas distintos y **asigna el rango 128–255** a caracteres específicos de cada alfabeto.

> Ejemplo:
> **ISO-8859-1 (Latin-1)** incluye las letras acentuadas y la **ñ**, usadas en Europa occidental.

> Otros ejemplos:
>
> - ISO-8859-2 → Europa Central (polaco, checo, húngaro)
> - ISO-8859-5 → Alfabeto cirílico (ruso)
> - ISO-8859-15 → Versión moderna de Latin-1 con el símbolo del euro (€)

💡 Relación con ASCII

- Los **128 primeros caracteres (0–127)** son **idénticos al ASCII original**.
- Los valores **128–255** se adaptan según la versión del estándar.

> ✅ En resumen:
> ISO-8859 fue la primera gran unificación internacional del ASCII extendido.

🔷 Unicode

📖 Definición

Unicode es un sistema universal de codificación de caracteres creado para **unificar todos los alfabetos y símbolos del mundo** bajo un mismo estándar.
Incluye desde letras latinas y cirílicas hasta ideogramas chinos, emojis y caracteres matemáticos.

- Cada carácter tiene un **código único (code point)**: `U+XXXX`
  (por ejemplo, `U+0041` = `A`, `U+00F1` = `ñ`)

> 🌍 Unicode es compatible con ASCII e ISO-8859, ya que conserva los mismos valores para los primeros 128 caracteres.

🧩 Formatos de codificación Unicode (UTF)

Unicode define los caracteres, pero **no cómo se guardan en memoria o en disco**.
Esa tarea la realizan los **formatos UTF (Unicode Transformation Format)**.

🔸 UTF-8

- ==Usa **entre 1 y 4 bytes** por carácter.
- Los caracteres ASCII (0–127) ocupan solo **1 byte**, manteniendo compatibilidad con ficheros antiguos.
- Es el formato **más usado actualmente** en Internet, Linux, bases de datos y Java.

> ✅ Ventaja: eficiente y compatible con ASCII.
> ❌ Inconveniente: los caracteres no latinos ocupan más espacio.

📘 Ejemplo:

```
A → 1 byte: 01000001
ñ → 2 bytes: 11000011 10110001
€ → 3 bytes: 11100010 10000010 10101100
```

🔸 UTF-16

- Usa **2 bytes por carácter** para la mayoría de los símbolos comunes,
  pero puede usar **4 bytes** para caracteres adicionales (como emojis o alfabetos asiáticos).
- Muy usado en sistemas Windows y en el interior de lenguajes como **Java** o **C#**.

> ✅ Ventaja: equilibrio entre espacio y compatibilidad internacional.
> ❌ Inconveniente: no compatible directamente con ASCII.

🔸 UTF-32

- Usa ==**4 bytes fijos por carácter**.
- Representa directamente el código Unicode sin compresión.

> ✅ Ventaja: acceso directo a cada carácter (simple y rápido).
> ❌ Inconveniente: **muy poco eficiente** en espacio (4 veces más que UTF-8 para texto ASCII).

⚙️ Comparativa general

Formato	Tamaño (bytes)	Compatible con ASCII	Uso principal
ASCII	1 (7 bits)	—	Lenguaje inglés básico
ISO-8859-1	1 (8 bits)	✅ Sí (0–127)	Idiomas europeos occidentales
UTF-8	1–4	✅ Sí	Web, Linux, Java, HTML
UTF-16	2–4	❌ No	Windows, Java interno
UTF-32	4	❌ No	Procesamiento interno de texto
